University of California, Berkeley

U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2013 *Paper* 310

Estimating Effects on Rare Outcomes: Knowledge is Power

Laura B. Balzer* Mark J. van der Laan†

*UC Berkeley, School of Public Health-Division of Biostatistics, lbbalzer@hsph.harvard.edu

†UC Berkeley, School of Public Health-Division of Biostatistics, laan@berkeley.edu

http://biostats.bepress.com/ucbbiostat/paper310

Estimating Effects on Rare Outcomes: Knowledge is Power

Laura B. Balzer and Mark J. van der Laan

Abstract

<blockquote>Many of the secondary outcomes in observational studies and randomized trials are rare. Methods for estimating causal effects and associations with rare outcomes, however, are limited, and this represents a missed opportunity for investigation. In this article, we construct a new targeted minimum loss-based estimator (TMLE) for the effect of an exposure or treatment on a rare outcome. We focus on the causal risk difference and statistical models incorporating bounds on the conditional risk of the outcome, given the exposure and covariates. By construction, the proposed estimator constrains the predicted outcomes to respect this model knowledge. Theoretically, this bounding provides stability and power to estimate the exposure effect. In finite sample simulations, the proposed estimator performed as well, if not better, than alternative estimators, including the propensity score matching estimator, inverse probability of treatment weighted (IPTW) estimator, augmented-IPTW and the standard TMLE algorithm. The new estimator remained unbiased if either the conditional mean outcome or the propensity score were consistently estimated. As a substitution estimator, TMLE guaranteed the point estimates were within the parameter range. Our results highlight the potential for double robust, semiparametric efficient estimation with rare events</blockquote>

1 Introduction

When the outcome of interest occurs infrequently, building prediction models and obtaining estimates of the intervention effect can be particularly challenging. For example, a recent study sought to examine the impact of planned place of delivery (obstetric unit or not) on perinatal mortality and neonatal morbidities, occurring in 250 of 63,827 births (0.39%) (Birthplace in England Collaborative Group, 2011). Due to the paucity of individual birth events, however, the researchers estimated the effect on a composite outcome measure. Likewise, tuberculosis is a main cause of mortality among HIV+ people (World Health Organization, 2013). Evaluating strategies to reduce its transmission are essential, but difficult due to the disease's relatively low incidence. Along the same lines, international consortiums have been established to investigate the burden and treatment for uncommon cancers (e.g. RARECARENet (2014)). While these outcomes are rare, a better understanding of their occurrence is likely to have important policy and health implications.

For binary outcomes or proportions, parametric logistic regression is often used to estimate the conditional odds ratio, given the exposure and measured covariates. Researchers, however, have recognized the inadequacy of this approach, when the outcome is rare (Concato et al., 1993; Harrell et al., 1996; Peduzzi et al., 1996; King and Zeng, 2001; Braitman and Rosenbaum, 2002; Cepeda et al., 2003). For example, simulations by Peduzzi et al. (1996) illustrated that estimates could be biased and inference unreliable if the number of outcomes per variable was less than 10. The authors also found problems with estimator convergence, statistical power and the validity of significance tests (i.e. type I error rates and confidence interval coverage). Harrell et al. (1996) cautioned against over-fitting and encouraged the use cross-validation or bootstrapping for model validation. Moreover, King and Zeng (2001) found that standard logistic models could substantially underestimate the probability of the outcome and offered a bias correction with accompanying software.

When dealing with rare events, several researchers have recommended estimators based on the propensity score, which is the conditional probability of being exposed, giving the covariates (Joffe and Rosenbaum, 1999; Braitman and Rosenbaum, 2002; Patorno et al., 2014). These methods avoid estimation of the conditional mean outcome and thereby are expected to perform well in the face of sparsity due to few events. Simulations by Cepeda et al. (2003) suggested that propensity score methods were less biased and more efficient than logistic regression for the mean outcome, when the number of events per coefficient was less than 8. The authors also cautioned that the performance of propensity score methods depended on the strength of the relationship between the covariates and the exposure.

Targeted minimum loss-based estimation (TMLE) is a general methodology for the construction of semiparametric, efficient substitution estimators (van der Laan and Rubin, 2006; van der Laan and Rose, 2011). The algorithm combines data-adaptive methods (i.e. machine learning techniques) with an additional updating step to reduce bias for the parameter of interest and to obtain valid statistical inference. TMLE has previously been applied in the context of rare outcomes in drug safety analyses (Lendle et al., 2013; Gruber and van der Laan, 2013). Building on the work of Gruber and van der Laan (2010), this paper proposes a new TMLE for statistical model $\mathcal{M}$, which bounds the conditional mean (risk) of the rare outcome from above by some small u.

The remainder of the article is organized as follows. In Section 2, the estimation problem and the theoretical motivation for the new TMLE are outlined. In particular, the statistical model, the target statistical parameter, the corresponding causal parameter and the efficient influence curve are discussed. Section 3 presents the rationale and procedure for the rare outcomes TMLE. Simulations are given in the Section 4. Sample `R` code is presented in the Appendix. The article concludes with a discussion of the advantages of the proposed TMLE and areas for future work.

2 The estimation problem

We are interested in estimating the impact of a binary exposure A on the risk of a rare outcome Y. For example, Y might be an indicator that the subject develops tuberculosis with an incidence rate of 255/100,000 per person-year in Sub-Saharan Africa (World Health Organization, 2013). Alternatively, Y might be the proportion of community members, who develop tuberculosis over the last year. Suppose we measure some baseline characteristics W that are predictors of both the exposure and outcome. In other words, W represents the set of measured confounders. Let $O = (W, A, Y)$ denote the observed data random variable with distribution P_0. Throughout, subscript 0 denotes the true, but unknown distribution. We first consider a nonparametric statistical model $\mathcal{M}_{np}$, which places no restrictions on the set of possible data distributions. We then focus on a semiparametric statistical model $\mathcal{M}$, which encodes bounds on the conditional risk of the rare outcome, given the exposure and covariates.

Throughout our goal is to estimate and obtain inference for the marginal risk difference:

$$\begin{aligned}\Psi(P_0) &= \sum_w \big[E_0(Y|A=1,W=w) - E_0(Y|A=0,W=w)\big] P_0(W=w) \\ &= E_0\big[\bar{Q}_0(1,W) - \bar{Q}_0(0,W)\big]\end{aligned}$$

where the summation generalizes to the integral for continuous covariates and $\bar{Q}_0(A,W) = E_0(Y|A,W)$ denotes the conditional mean outcome, given the exposure and covariates. This estimand is the difference in the strata-specific risk of the outcome, averaged (standardized) with respect to the baseline covariate distribution.

Under the necessary causal assumptions, this statistical parameter equals the causal risk difference and the average treatment effect: $\Psi^F(P_X) = E(Y_1) - E(Y_0)$. Here Y_a represents the counterfactual outcome, if possibly contrary-to-fact, the unit received exposure $A = a$. Briefly, the observed data can be considered as a time-ordered missing data structure on the full data $X^F = (W, Y_1, Y_0) \sim P_X$, with censoring variable A (Neyman, 1923; Rubin, 1974). Alternatively, the observed data can be considered as a draw from a distribution described by a structural causal model, which also generates the counterfactual outcomes Y_a (Pearl, 2000). To express the target causal parameter $\Psi^F(P_X)$ as a function of the observed data distribution P_0, we need two assumptions. First, there must be no unmeasured confounders of the effect of the exposure on the outcome. Secondly, there must be sufficient variability in the treatment assignment. In other words, the propensity score $P_0(A = 1|W)$ must be bounded away from 0 and 1. This condition is known as the positivity assumption. The reader is referred to Petersen and van der Laan (2014) for an introduction to the causal framework and discussion of the necessary assumptions.

The challenge, addressed in this paper, is estimation with rare outcomes. This challenge is illuminated by studying the efficient influence curve (function) of the target parameter Ψ at a probability distribution P in a nonparametric statistical model $\mathcal{M}_{np}$ (Bickel et al., 1993; van der Laan and Rose, 2011):

$$D^*(P)(O) = \left(\frac{\mathbb{I}(A=1)}{P(A=1|W)} - \frac{\mathbb{I}(A=0)}{P(A=0|W)}\right)\big(Y - \bar{Q}(A,W)\big) + \bar{Q}(1,W) - \bar{Q}(0,W) - \Psi(P)$$

where $\mathbb{I}(\cdot)$ is the indicator function. $D^*(P)(O)$ is the critical ingredient in the construction of double-robust, semiparametric, efficient estimators (van der Laan and Robins, 2003; van der Laan and Rose, 2011). Specifically, $D^*(P)(O)$ can be represented as an estimating function in $\Psi(P)$ as in augmented inverse probability of treatment weighting (A-IPTW) or used to fluctuate initial estimates in the substitution estimator TMLE.

The information for learning the target parameter is captured by the sample size n divided by the variance of the efficient influence curve at the true distribution P_0. Sparsity or low information can occur when the propensity score $P_0(A = 1|W)$ approaches 0 or 1 for certain treatment-covariate combinations. The impact of violations of the positivity assumption on estimator performance are demonstrated in Petersen et al. (2012). Sparsity can also occur when there are very few events and the conditional mean $\bar{Q}_0(A, W)$ is or approaches 1 for some treatment-covariate combinations. In either case, the variance of the efficient influence curve can become large, and the sample size needed to detect effects with sufficient power correspondingly large.

This paper considers the semiparametric statistical model $\mathcal{M}$, which incorporates knowledge that the conditional mean outcome $\bar{Q}(A, W)$ is bounded from above by some small constant u. For rare binary outcomes, this knowledge has been noted by other researchers (Beck et al., 2000; King and Zeng, 2001). Consider, for example, an unusual outcome with incidence 0.5%. With this rare of an event, it seems unlikely that the conditional mean outcome exceeds 10% for any combination of the exposure and measured covariates. Instead, the researcher might have knowledge that this conditional mean does not exceed 5% for the study population. Formally, the statistical model $\mathcal{M}$ is the set of possible data generating distributions:

$$\mathcal{M} = \left\{P : \bar{Q}(A, W) \in [\ell, u]\right\}$$

for some bounds $0 \leq \ell < u < 1$. The specification of the bounds can be based on subject matter knowledge or selected with cross-validation, as discussed in Section 3.3. We place no restrictions on the marginal distribution of baseline covariates $P(W)$ or on the propensity score $P(A = 1|W)$.

The global constraints in the semiparametric statistical model $\mathcal{M}$ do not change the efficient influence curve. They do, however, restrict its variance and thereby improve the information for estimating the target parameter. To see this, let us assume the propensity score is bounded away from 0 and 1. In other words, we assume there are no positivity violations. Further, let us set the lower bound ℓ to 0. Then we can express the conditional mean outcome as $\bar{Q}(A, W) = u\tilde{Q}(A, W)$ for some function $\tilde{Q}(A, W) \in [0, 1]$. For a binary outcome Y, the variance of the efficient influence curve at some P in the semiparametric model $\mathcal{M}$ can be rewritten as

$$Var\left[D^*(P)(O)\right] = uE\left[\frac{\tilde{Q}(1, W)\left(1 - u\tilde{Q}(1, W)\right)}{P(A = 1|W)} + \frac{\tilde{Q}(0, W)\left(1 - u\tilde{Q}(0, W)\right)}{P(A = 0|W)}\right] + u^2 E\left[\left(\tilde{Q}(1, W) - \tilde{Q}(0, W) - \tilde{\psi}\right)^2\right]$$

where $\tilde{\psi} = E[\tilde{Q}(1, W) - \tilde{Q}(0, W)]$. See Appendix A for the accompanying proof. As a result, the variance of the efficient influence curve can be expressed as the upper limit u times a bounded function of means. In other words, the variance of the efficient influence curve is dampened by a factor of u, and the corresponding information for learning the target parameter from sample size n is amplified by a factor of $1/u$.

As originally noted in van der Laan (2008) and elaborated in Appendix B, this provides motivation for the construction of an efficient, substitution estimator, which is expected to have reasonable power in finite samples. Briefly, the variance of the efficient influence curve establishes a lower bound on the asymptotic variance of all regular, asymptotically linear estimators (Bickel et al., 1993). A TMLE, enforcing the model constraints, is expected to achieve or nearly achieve this efficiency bound even when the estimator of $\bar{Q}_0(A, W)$ converges to a misspecified limit (Proof in Appendix B). Intuitively, bounding the estimates of $\bar{Q}_0(A, W)$ to be $\leq u$ limits the potential deviations between the estimates and the truth, and since $\bar{Q}_0(A, W)$ yields the optimal variance, this helps us to get closer to the efficiency bound. Thus, incorporating model knowledge is expected

to decrease the estimator's asymptotic variance as well as improve stability in finite samples. This provides asymptotic as well as finite sample motivations for the development of a new TMLE for the previously unconsidered statistical model $\mathcal{M}$.

3 Estimation

Suppose we have n observations $O = (W, A, Y)$ drawn independently from P_0. A simple substitution estimator (i.e. the G-Computation estimator (Robins, 1986)) for the marginal risk difference $\Psi(P_0)$ can be implemented with the following steps. First, we obtain an estimate of the conditional mean outcome $\bar{Q}_0(A, W)$. Let us denote an initial estimator based on n observations as $\bar{Q}_n(A, W)$. Then we obtain the predicted outcomes for each observation under the exposure $\bar{Q}_n(1, W)$ and under the control $\bar{Q}_n(0, W)$. The sample average difference in these predicted outcomes provides a point estimate for $\Psi(P_0)$. The final step corresponds to estimating the marginal covariate distribution with the empirical proportion.

In some cases, we can estimate $\bar{Q}_0(A, W)$ with sample proportion in each exposure-covariate strata. This non-parametric estimator can quickly become ill-defined when there are many exposure-covariate combinations and can suffer from over-fitting, especially with rare outcomes. In some cases, we may have the background knowledge to support parametric regression for $\bar{Q}_0(A, W)$. We could, for example, run main terms logistic regression of the outcome Y on the exposure A and measured covariates W. If this assumed parametric model is wrong, however, we are in danger of biased point estimates and inference. Even if the assumed parametric model is correct, we are in danger of biased point estimates and inference due to over-fitting.

Data-adaptive algorithms, using cross-validation (i.e. sample splitting), can help minimize these dangers. Super Learner, for example, uses cross-validation to select the algorithm with best performance or to build the optimal convex combination of candidate algorithms (van der Laan et al., 2007). When the outcome is binary or a proportion, performance of candidate estimators can be evaluated with the negative log-likelihood loss function. As before, a point estimate can be obtained by plugging in the predicted outcomes, setting $A = 1$ and $A = 0$, to the parameter mapping. Inference, however, must respect the model building process. (Treating the final regression model as if it were *a priori*-specified can result in misleading inference.) Furthermore, estimating the conditional expectation $\bar{Q}_0(A, W)$ is a more ambitious task than estimating a single number $\Psi(P_0) \in \mathbb{R}$. Thereby, an estimator of $\bar{Q}_0(A, W)$ will have the wrong bias-variance trade-off for the parameter of interest $\Psi(P_0)$.

TMLE provides a solution to several of these challenges. In TMLE, the initial estimator of $\bar{Q}_0(A, W)$ is updated to reduce bias for $\Psi(P_0)$ and to attain valid statistical inference. This updating is accomplished with a loss function and a carefully selected submodel. For a binary or bounded continuous outcome, the negative log-likelihood loss function and the following logistic regression model can be used to update the initial estimator (Gruber and van der Laan, 2010):

$$logit\big[\bar{Q}_n(A, W)(\epsilon)\big] = logit\big[\bar{Q}_n(A, W)\big] + \epsilon H_n(A, W)$$
$$\text{with covariate } H_n(A, W) = \left(\frac{\mathbb{I}(A = 1)}{P_n(A = 1|W)} - \frac{\mathbb{I}(A = 0)}{P_n(A = 0|W)}\right)$$

and with ϵ as the univariate parameter. *Logit* refers to the logarithm of the odds and $P_n(A = 1|W)$ denotes an estimate of the propensity score. This combination of loss function and fluctuation model returns the initial estimator $\bar{Q}_n(A, W)$ at $\epsilon = 0$ and has score equal to the relevant component of the efficient influence curve at $\epsilon = 0$. To estimate the fluctuation parameter ϵ, we run logistic regression of the outcome Y on the covariate $H_n(A, W)$ with the *logit* of the initial estimates as

offset. Plugging in the estimated coefficient ϵ_n yields the targeted updates. A point estimate can then be obtained by substituting the targeted estimates into the parameter mapping.

The standard TMLE algorithm for a binary or bounded continuous outcome is not optimal for our semiparametric statistical model $\mathcal{M}$. Specifically, this TMLE does not enforce the global constraints on the conditional mean outcome. First, initial estimates $\bar{Q}_n(A,W)$ based on the log-likelihood loss function are guaranteed to be within $(0,1)$, but are possibly outside of the model bounds $[\ell,u]$. Secondly, the logistic regression, used to update the initial estimator, is technically not a submodel, because it does not respect the constraints implied by $\mathcal{M}$. Therefore, the updated estimates $\bar{Q}_n^*(A,W)$ are guaranteed to be between $(0,1)$, but may be outside of the model bounds $[\ell,u]$. As a result, the asymptotic and finite sample performance of a TMLE, using this loss function and parametric working model, is expected to sub-optimal. The algorithm is expected to be unstable due to sparsity from rare events.

3.1 The rare outcomes TMLE

To incorporate the model knowledge, we propose a linear transformation of the rare outcome $Y \in [0,1]$ by subtracting off the lower bound ℓ and dividing by the deviation between the upper u and lower bounds:

$$\tilde{Y} = \frac{Y-\ell}{u-\ell} \in \left[\frac{-\ell}{u-\ell'}, \frac{1-\ell}{u-\ell}\right]$$

Suppose, for example, the bounds on $\bar{Q}_0(A,W)$ are $[0, 0.05]$. Then the transformed outcome $\tilde{Y}$ would be bounded between 0 and 20. The mapping between the conditional mean of the original outcome Y and the conditional mean of the transformed outcome $\tilde{Y}$ is given by

$$\bar{Q}(A,W) = \ell + (u-\ell)\tilde{Q}(A,W)$$

where $\tilde{Q}(A,W)$ is a bounded function of exposure and covariates (A,W). The analogous transformation was proposed by Gruber and van der Laan (2010) for the TMLE of a bounded continuous outcome $Y \in [a,b]$.

The negative quasi-log-likelihood is a valid loss function for initial estimation and targeting of the transformed mean $\tilde{Q}_0(A,W)$:

$$-\mathcal{L}(\tilde{Q})(O) = log\left[\tilde{Q}(A,W)^{\tilde{Y}}\big(1-\tilde{Q}(A,W)\big)^{1-\tilde{Y}}\right]$$

To update an initial estimator of $\tilde{Q}_n(A,W)$, we can use the logistic fluctuation submodel: $logit\big[\tilde{Q}_n(A,W)(\epsilon)\big] = logit\big[\tilde{Q}_n(A,W)\big] + \epsilon H_n(A,W)$ with covariate $H_n(A,W)$ defined as above. This combination of loss function and fluctuation submodel will generate a score proportional to the relevant component of the efficient influence curve at zero fluctuation:

$$\begin{aligned}\frac{d}{d\epsilon}\mathcal{L}(\tilde{Q}_n(\epsilon))(O)\Big|_{\epsilon=0} &= H_n(A,W)\big(\tilde{Y}-\tilde{Q}_n(A,W)\big)\\ &= H_n(A,W)\left(\frac{Y-\ell}{u-\ell}-\tilde{Q}_n(A,W)\right)\\ &= \frac{1}{u-\ell}H_n(A,W)(Y-\bar{Q}_n(A,W))\end{aligned}$$

Through this transformation, the initial and targeted estimates are guaranteed to satisfy the model constraints. This will provide robustness. The targeted estimates can then be rescaled and substituted in the parameter mapping. The proposed loss function and parametric submodel define

a new TMLE of the target parameter $\Psi(P_0)$ in the semiparametric statistical model $\mathcal{M}$, which encodes bounds on the conditional risk of the rare outcome $\bar{Q}_0(A,W) \in [\ell, u]$. Hereafter, we refer to the proposed estimator as the rare outcomes TMLE.

3.2 Step-by-step implementation

Step 1: Transform the outcome. We first transform the outcome Y into $\tilde{Y}$ by subtracting the lower bound ℓ and dividing by the difference between the upper and lower bounds $(u - \ell)$.

Step 2: Estimate the transformed mean. An initial estimate of the conditional mean of the transformed outcome $\tilde{Q}_0(A,W)$ can be based on logistic regression of $\tilde{Y}$ on the exposure A and baseline covariates W. Since the outcome is no longer between 0 and 1, standard software may yield error messages. Example `R` code, using the `optim` function, is given in the Appendix (R Core Team, 2014). More data-adaptive methods, such as Super Learner, can be implemented as long as the library of algorithms respect the statistical model (van der Laan et al., 2007).

Step 3: Estimate the propensity score. An initial estimate of the conditional probability of being exposed given the covariates $P_0(A = 1|W)$ is also required, as it makes up the covariate $H_n(A,W)$ in the fluctuation submodel. Using the log-likelihood loss function, the propensity score could be estimated with parametric logistic regression or with more data-adaptive methods.

Step 4: Target the initial estimator. Run logistic regression of the transformed outcome $\tilde{Y}$ on the covariate $H_n(A,W)$ with offset as the *logit* of the initial estimates $\tilde{Q}_n(A,W)$. The estimated coefficient ϵ_n is then plugged into yield the updates:

$$\tilde{Q}_n^*(A,W) = \tilde{Q}_n(A,W)(\epsilon_n) = expit\big[logit[\tilde{Q}_n(A,W)] + \epsilon_n H(A,W)\big]$$

where *expit* is the inverse-*logit*. The process of estimating the fluctuation parameter ϵ and updating is iterated until convergence, which occurs here in a single step.

Step 5: Transform and plug-in the targeted estimates. The targeted estimate of the conditional mean of the transformed outcome, denoted $\tilde{Q}_n^*(A,W)$, can be mapped into a targeted estimate of the conditional mean of the original outcome by

$$\bar{Q}_n^*(A,W) = \ell + (u - \ell)\tilde{Q}_n^*(A,W)$$

We obtain a point estimate by substituting in the targeted estimates of the conditional mean under the exposure $\bar{Q}_n^*(1,W)$ and under no exposure $\bar{Q}_n^*(0,W)$ into the parameter mapping:

$$\Psi_{n,rtmle}(P_n) = \frac{1}{n}\sum_{i=1}^{n}\left(\bar{Q}_n^*(1,W_i) - \bar{Q}_n^*(0,W_i)\right)$$

Step 6: Obtain inference. Under regularity conditions, $\Psi_{n,rtmle}(P_n)$ is an asymptotically linear estimator of $\Psi(P_0)$ (van der Laan and Rose, 2011). Its limit distribution is normal with mean 0 and variance given by the variance of its influence curve divided by sample size n. Therefore, 95% confidence intervals can be constructed as the point estimate $\psi_n^* \pm 1.96\sigma_n/\sqrt{n}$, where σ_n^2 is the sample variance of the estimated influence curve:

$$IC_n(O) \quad = \left(\tfrac{\mathbb{I}(A=1)}{P_n(A=1|W)} - \tfrac{\mathbb{I}(A=0)}{P_n(A=0|W)}\right)\left(Y - \bar{Q}_n^*(A,W)\right) + \bar{Q}_n^*(1,W) - \bar{Q}_n^*(0,W) - \psi_n^*$$

where ψ_n^* denotes the resulting point estimate. Alternative approaches for variance estimation include the non-parametric bootstrap or a substitution estimator for the variance. The former might be problematic with rare binary outcomes as some bootstrapped samples may not have any

events. The latter would guarantee the bounds on the variance are respected and is an area of future work.

By construction, the rare outcomes TMLE solves the efficient score equation. Specifically, the empirical mean of the efficient influence curve at the targeted estimator $\bar{Q}^*(A,W)$ and initial estimator $P_n(A|W)$ is zero. As a result, the TMLE is double robust: ψ_n^* will be a consistent estimator for $\psi_0 = \Psi(P_0)$ if either the conditional mean function $\bar{Q}_0(A,W)$ or the propensity score $P_0(A=1|W)$ is consistently estimated. As shown below, this property translates into important bias gains in finite samples. If both functions are consistently estimated and the propensity score satisfies the positivity assumption, the proposed TMLE will be asymptotically efficient in that its influence curve equals the efficient influence curve. As illustrated below, this property translates into important variance and power gains in finite samples.

3.3 Selecting the upper bound u with cross-validation

Thus far, we have assumed that the upper bound u is known. (The lower bound ℓ can trivially be set to 0.) Such knowledge, however, may be unavailable in all applications. When the specified upper bound is larger than needed, the gains in estimator performance will be attenuated. In the extreme when $u = 1$, the rare outcomes TMLE will reduce to the standard TMLE algorithm for binary or bounded continuous outcomes. If the specified upper bound is too small, then the targeted estimator of $\bar{Q}_0(A,W)$ will be inconsistent. Nonetheless, if the propensity score is known or consistently estimated, then the target parameter ψ_0 will still be consistently estimated due to the double robustness property. Moreover, the targeted estimates $\bar{Q}_n^*(A,W)$ will still be bounded from above by u, which translates into important variance gains. Thereby, cross-validation can be used to select u, when such knowledge is not available *a priori*. Specifically, the upper bound can be selected by minimizing the cross-validated risk of candidate estimators $\bar{Q}_{n,u}(A,W)$, which are now indexed by an upper bound u. The previously stated properties (e.g. double robustness, asymptotic linearity and efficiency) should hold when the upper bound u is selected data-adaptively. We implement this cross-validation selector in the following simulations.

4 Simulations

The following simulation studies compare the finite sample performance of the standard TMLE with the proposed rare outcomes TMLE (rTMLE). For comparison, we also include the propensity score matching (PSM) estimator (Rosenbaum and Rubin, 1983), inverse probability of treatment weighted (IPTW) estimator (Hernán et al., 2000) and augmented inverse probability of treatment weighted (AIPTW) estimator (Robins, 2000; van der Laan and Robins, 2003). The first two methods rely solely on estimation of the propensity score and may have superior performance under sparsity due to rare outcomes. AIPTW requires estimation of both the conditional mean outcome and the propensity score, but is double robust and asymptotically efficient under consistent estimation of both functions. AIPTW is an estimating equation (i.e. not a substitution estimator) and thereby can result in impossible parameter estimates in the context of sparsity (e.g. probabilities less than 0 or greater than 1 (Lendle et al., 2013)).

For the PSM estimator, we used the `Matching` package in `R` for 1:1 matching based on the estimated propensity score and calculated the point estimate as

$$\psi_{n,PSM} = \frac{1}{n}\sum_{i=1}^{n} \hat{Y}_i(1) - \hat{Y}_i(0), \text{ where } \hat{Y}_i(a) = \begin{cases} Y_i & \text{if } A_i = a \\ Y_{M_i} & \text{if } A_i \neq a \end{cases}$$

with Y_{M_i} denoting the outcome of observation matched to unit i based on the estimated propensity scores (Sekhon, 2011). A point estimate from IPTW is given by the following weighted mean

$$\psi_{n,IPTW} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\mathbb{I}(A_i=1)}{P_n(A_i=1|W_i)} - \frac{\mathbb{I}(A_i=0)}{P_n(A=0|W_i)}\right)Y_i$$

A point estimate from AIPTW is attained by directly solving the efficient score equation:

$$\begin{aligned}\psi_{n,AIPTW} = \ & \tfrac{1}{n}\textstyle\sum_{i=1}^{n}\Bigg[\left(\frac{\mathbb{I}(A_i=1)}{P_n(A_i=1|W)} - \frac{\mathbb{I}(A_i=0)}{P_n(A_i=0|W_i)}\right)\left(Y_i - \bar{Q}_n(A_i,W_i)\right) \\ & +\bar{Q}_n(1,W_i) - \bar{Q}_n(0,W_i)\Bigg]\end{aligned}$$

where $\bar{Q}_n(A,W)$ denotes a non-targeted estimate of the conditional mean outcome. Inference was based on the Abadie-Imbens standard error estimator for the PSM estimator (Abadie and Imbens, 2006; Sekhon, 2011) and the estimated influence curve for the others. Wald-type confidence intervals were calculated as $\psi_n \pm 1.96se_n$, where ψ_n denotes the point estimate and se_n denotes the estimated standard error. Likewise, tests of the null hypothesis of no effect were based on test statistic $T_n = \psi_n/se_n$.

4.1 Simulation 1: Individual-level data

The finite sample performance of the estimators was assessed by drawing 2000 samples of sizes $n = 1000$ and $n = 2500$ according to the following process. First, we generated three baseline covariates:

$$W1 \sim Normal(0, 0.25^2), \quad W2 \sim Uniform(0,1), \quad W3 \sim Bernoulli(0.5)$$

The exposure A was drawn from a Bernoulli distribution with probability

$$P_0(A=1|W) = expit(-0.5 + W1 + W2 + W3)$$

With this exposure mechanism, there were no positivity violations; the propensity score was bounded between 20% and 92%. Finally, the binary outcome was drawn from a Bernoulli distribution with probability

$$P_0(Y=1|A,W) = expit(-3 + 2^*A + W1 + 2^*W2 - 4^*W3 + 0.5^*A^*W1)/15$$

The resulting marginal probability of the outcome was $E_0(Y) = 1.1\%$. The true bounds on the conditional mean $\bar{Q}_0(A,W)$ were $[0\%, 6.2\%]$, and the true value of the statistical parameter was ψ_0=1.3%.

The propensity score $P_0(A = 1|W)$ was estimated with the correctly specified main terms logistic model and a misspecified regression model, failing to adjust for $W3$. The conditional mean outcome $\bar{Q}_0(A,W)$ was estimated with the correctly specified logistic regression model as well as a misspecified regression, only adjusting for $W1$. For the rare outcomes TMLE, the lower bound was set 0 and the upper bound was selected data-adaptively from the set {2.5%, 5%, 7.5%, 10%} using cross-validation with the log-likelihood loss function.

The results for this simulation are presented in Figures 1 and 2. As expected, the PSM estimator and IPTW exhibited low bias when the regression model for the propensity score $P_0(A = 1|W)$ was

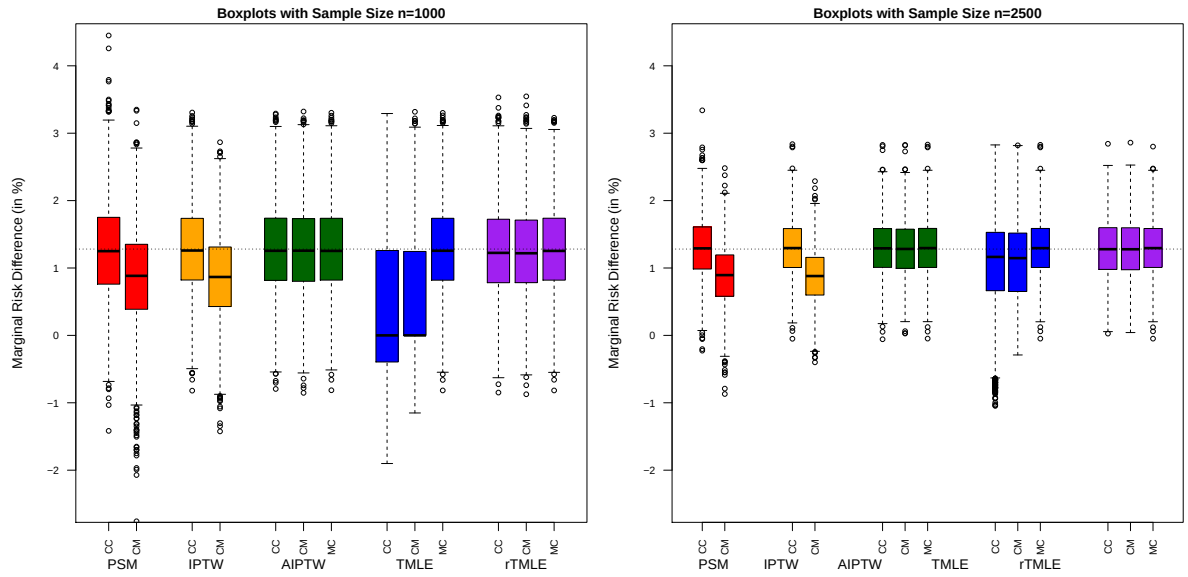

Figure 1: Box plots of the point estimates from various algorithms for sample sizes $n = 1000$ (left) and $n = 2500$ (right) for Simulation 1. The x-axis denotes the estimator and the regression specification. CC indicates both the outcome regression and the propensity score are correctly specified. CM indicates the outcome regression is correctly specified, but propensity score misspecified. MC indicates the outcome regression is misspecified, but the propensity score correctly specified. The dashed line indicates the true value $\psi_0 = 1.3\%$.

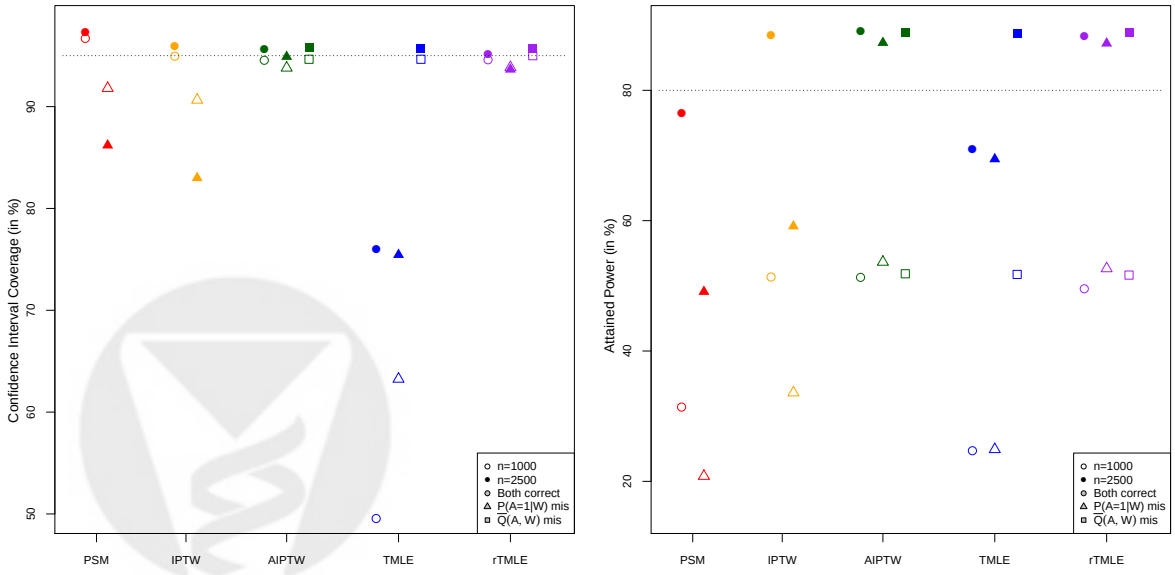

Figure 2: Attained confidence interval coverage and power for the estimators in Simulation 1. The unfilled and filled points denote sample sizes of $n = 1000$ and $n = 2500$, respectively. Circles indicate both the outcome regression and the propensity score are correctly specified. Triangles indicate the outcome regression is correctly specified, but propensity score misspecified. Squares indicate the outcome regression is misspecified, and propensity score correctly specified. The dashed lines indicate 95% confidence interval coverage and 80% power, respectively.

correctly specified, but were biased otherwise. This bias did not disappear with sample size. Likewise, these estimators had poor confidence interval coverage under misspecification of the propensity score. Even under correct specification of this function, the PSM estimator had notably lower power than IPTW, AIPTW or rTMLE. Indeed, neither the PSM estimator nor IPTW are efficient estimators.

In theory, the other estimators solve the efficient score equation directly (AIPTW) or during the targeting step (TMLE, rTMLE). As a result, these estimators are double robust. For these simulations, however, the standard TMLE exhibited substantial bias when estimating $\bar{Q}_0(A, W)$ with the correctly specified regression model, which included 4 main terms plus an interaction (Figure 1). Due to the paucity of events, logistic regression of the untransformed outcome Y was unstable. Thereby, the updating step, which involved fitting an additional coefficient ϵ, did little to reduce bias. As a result, its confidence interval coverage was poor under correct specification of the outcome regression (Figure 2). On the other hand, when the initial estimator of $\bar{Q}_0(A, W)$ was based on the misspecified regression, the performance of the standard TMLE was comparable to AIPTW and rTMLE. Recall the misspecified regression model for $\bar{Q}_0(A, W)$ only included main terms for the exposure and $W1$. The smaller adjustment set lead to more stable initial estimates, but also complete reliance on consistent estimation of the propensity score for confounding control and consistency. Thereby, a potential approach when estimating effects with rare outcomes is to use a smaller adjustment set for $\bar{Q}_0(A, W)$ and thereby sacrifice double robustness and efficiency for greater stability and potentially reduced bias. This approach was taken by Gruber and van der Laan (2013).

The performance of AIPTW and rTMLE did not suffer when fitting $\bar{Q}_0(A, W)$ in the larger, correctly specified regression model. As shown in Figure 1, both AIPTW and rTMLE had low bias when either the outcome regression, the propensity score or both were correctly specified. Furthermore, these estimators had good confidence interval coverage at both sample sizes and attained power for the larger sample (Figure 2). Therefore, it seemed that the performance of AIPTW and rTMLE were comparable for this set of simulations. It is worth emphasizing, however, that rTMLE is a substitution estimator and thereby guaranteed to respect the global constraints in the model. In particular, AIPTW yielded negative (impossible) estimates for the marginal risk under no exposure (Table 1).

	$n = 1000$	$n = 2500$
Both Correct	120	5
$P_0(A = 1\|W)$ mis	260	11
$\bar{Q}_0(A, W)$ mis	130	6

Table 1: For Simulation 1, the number of negative estimates for the marginal risk under no exposure from AIPTW with samples of size $n = 1000$ and $n = 2500$.

Under the null, the PSM estimator and IPTW had good type I error control when the propensity was correctly specified (Results not shown). Under misspecification, their type I error rates exceeded 20%. The standard TMLE also suffered from high type I error rates, when both regression models were correctly specified. AIPTW and rTMLE maintained nominal type I error rates for both sample sizes and all regression specifications.

4.2 Simulation 2: Group-level data

For this set of simulations, we focused on clustered data, where the covariates, exposure and outcome were measured or aggregated to the cluster-level. For 2,000 simulations of $n = 100$ units, three

baseline covariates $\big(W1, W2, W3\big)$ and the exposure A were generated according to the above process. We also simulated two additional covariates $W4$ and $W5$ as independent draws from a normal distribution with mean 0 and standard deviation 0.25. The cluster-level outcome Y was the empirical mean of 2,500 independent Bernoulli's with a cluster-specific risk of

$$P_0(Y = 1|A, W) = expit(-2.5 + 0.5^*A + W1 + 2^*W2 - 2^*W3)/15$$

The average outcome across the clusters was $E_0(Y) = 0.99\%$. The true bounds on the conditional mean $\bar{Q}_0(A, W)$ were $[0\%, 4.7\%]$, and the true value of the statistical parameter was ψ_0=0.36%. For completeness, the null scenario was also generated by simulating the outcomes as if all clusters were exposed.

As suggested by the previous simulations, a possible approach for estimation with rare outcomes is using a small adjustment set for $\bar{Q}_0(A, W)$ and fully relying on consistent estimation of the propensity score $P_0(A = 1|W)$. To evaluate this approach, we estimated the conditional mean outcome with the unadjusted treatment-specific mean $\bar{Q}_n(A)$. For comparison, we also employed Super Learner for initial estimation of transformed mean $\tilde{Q}_0(A, W)$ in rTMLE. For simplicity, we limited our library of algorithms to logistic regressions, building up from a single main term to four terms (the exposure plus three covariates). For example, a candidate algorithm included as main terms the exposure A, the baseline confounder $W1$ as well as the two irrelevant covariates $W4$ and $W5$. This simple library was selected for illustration; in practice, the inclusion of more flexible algorithms is recommended. The bounds for rTMLE were selected from the set {2.5%, 5%, 7.5%, 10%} with cross-validation and the log-likelihood loss. The propensity score was estimated according to the correctly specified logistic regression model.

The simulation results are given in Table 2. Since the estimated propensity score was based on the correct regression model, all algorithms were unbiased. When using an unadjusted (initial) estimator $\bar{Q}_n(A)$, the double robust estimators did not substantially outperform the PSM estimator. Instead, AIPTW, TMLE and rTMLE were under-powered, and there was evidence of conservative inference, as indicated by over-coverage of confidence intervals and lower than nominal Type I error rates. Furthermore, when there was a treatment effect, the mean squared error (MSE) of standard TMLE, using the unadjusted initial estimate $\bar{Q}_n(A)$, was notably larger than that of the other estimators. This suggests that even when a small adjustment set is used, rTMLE can provide stability over the standard algorithm. The rTMLE, using Super Learner for initial estimation, had good confidence interval coverage and type I error control. Furthermore, when there was a treatment effect, the rTMLE, using Super Learner, attained highest power of 95%. This demonstrates the potential gains with data-adaptive estimation in the rTMLE algorithm.

5 Discussion

In this paper, we proposed a new TMLE for evaluating causal effects and estimating associations with rare outcomes. The efficient influence curve (function) provided theoretical motivation for the new estimator. Specifically, sparsity or low information can arise when the propensity score approaches 0 or 1 for some exposure-covariate combinations or when outcome is rare and the conditional mean outcome approaches 1 for some exposure-covariate combinations. Thereby, the information for estimating the impact on a rare outcome is substantially improved when the conditional mean outcome is bounded from above by some small u. These bounds can be based on subject matter knowledge or selected with cross-validation and are encoded in a semiparametric statistical model. Estimators, incorporating this model knowledge, are expected to be less variable asymptotically and more stable in finite samples. Furthermore, if the estimator of conditional mean

	Treatment effect: ψ_0=0.36%				Null: $\psi_0 = 0\%$			
	ψ_n (%)	MSE[a]	Cov.[b]	Power	ψ_n (%)	MSE[a]	Cov.[b]	α
PSM	0.36	1.4E-6	1.00	0.53	0.00	1.9E-6	0.99	0.01
IPTW	0.35	6.3E-7	1.00	0.02	0.00	7.6E-7	1.00	0.00
AIPTW	0.35	7.4E-7	1.00	0.47	0.00	9.1E-7	1.00	0.00
TMLE	0.38	1.0E-4	1.00	0.55	0.02	1.0E-4	1.00	0.00
rTMLE	0.35	7.2E-7	1.00	0.55	0.00	8.6E-7	1.00	0.00
rTMLE$_{SL}$	0.36	7.3E-7	0.96	0.95	0.00	1.1E-6	0.97	0.03

[a] mean squared error (bias-squared + variance)

[b] confidence interval coverage

Table 2: Estimator performance over 2,000 simulations of $n = 100$ clusters in Simulation 2. The rows indicate the estimator with "SL" denoting that Super Learner was used for initial estimation of $\tilde{Q}_0(A, W)$.

$\bar{Q}_0(A, W)$ converges to a misspecified limit but still satisfies the model bounds, the corresponding TMLE is expected to achieve or nearly achieve the efficiency bound.

In finite sample simulations, the proposed rare outcomes TMLE performed as well or outperformed the alternative estimators. The PSM estimator and IPTW were biased under misspecification of the propensity score. The standard TMLE algorithm suffered from bias and poor confidence interval coverage when the adjustment set for the conditional mean outcome was large. Both AIPTW and rTMLE were robust to model misspecification and remained unbiased if either the conditional mean outcome or the propensity score were consistently estimated. AIPTW, however, is not a substitution estimator and yielded negative (impossible) risk estimates. In contrast, the proposed TMLE respected the global knowledge in the statistical model. Our simulations further highlighted the potential for data-adaptive estimation to avoid parametric assumptions and to increase power.

The proposed TMLE is generalizable for estimation of other parameters, including the risk ratios, odds ratios and the impacts of longitudinal exposures. The TMLE is also applicable to other sampling designs. Specifically, case-control studies are commonly employed to increase robustness and efficiency of the analysis of rare events. There are several well established methods that correct for selection on the outcome (Anderson, 1972; Prentice and Breslow, 1978; King and Zeng, 2001; Robins, 1999; Mansson et al., 2007). For example, van der Laan (2008) presented a general mapping of loss functions, substitution estimators and estimating equations developed for prospective sampling (i.e. cohort sampling) into loss functions, substitution estimators and estimating equations for biased sampling (e.g. case-control sampling). The estimator's properties, such as double robustness and asymptotic efficiency, are preserved under the mapping. As noted by van der Laan (2008), however, the sample size needed to detect effects on the order of the outcome prevalence will be very large, unless the conditional probability of the outcome is bounded from above by some small constant u. The estimator, proposed in this article, satisfies this condition by construction and thereby has the potential to achieve higher power than its unconstrained counterpart in finite samples. In other words, the mapping provided in van der Laan (2008) will allow us to weight rTMLE appropriately for case-control sampling. We expect the resulting estimator to offer an improvement in terms of stability and power in finite samples.

Appendix

Appendix A: Variance of the efficient influence curve at $P_0 \in \mathcal{M}$

Suppose the outcome Y is binary, and consider the semiparametric statistical model $\mathcal{M}$, bounding the conditional probability of the outcome $\bar{Q}(A,W)$ by some $[\ell, u]$ with $0 \leq \ell < u < 1$. For simplicity, let us us assume the lower bound ℓ is 0 and re-express the conditional mean function as

$$\bar{Q}(A,W) = u\tilde{Q}(A,W)$$

for some $\tilde{Q}(A,W) \in [0,1]$. Let $g_0(A|W)$ denote the conditional distribution of the exposure, given the covariates $P_0(A|W)$. Then the variance of the efficient influence curve at the true but unknown distribution P_0 can be written as

$$\begin{aligned}
&Var[D^*(Q_0, g_0)] \\
&= E_0\left[\left(\frac{I(A=1)}{g_0(1|W)} - \frac{I(A=0)}{g_0(0|W)}\right)^2 \left(Y - \bar{Q}_0(A,W)\right)^2\right] + E_0\left[\left(\bar{Q}_0(1,W) - \bar{Q}_0(0,W) - \Psi(Q_0)\right)^2\right] \\
&= E_0\left[\left(\frac{I(A=1)}{g_0(1|W)^2} + \frac{I(A=0)}{g_0(0|W)^2}\right) E_0\left[\left(Y - \bar{Q}_0(A,W)\right)^2 \middle| A, W\right]\right] + E_0\left[\left(\bar{Q}_0(1,W) - \bar{Q}_0(0,W) - \Psi(Q_0)\right)^2\right] \\
&= E_0\left[\frac{\bar{Q}_0(1,W)\left(1 - \bar{Q}_0(1,W)\right)}{g_0(1|W)} + \frac{\bar{Q}_0(0,W)\left(1 - \bar{Q}_0(0,W)\right)}{g_0(0|W)}\right] + E_0\left[\left(\bar{Q}_0(1,W) - \bar{Q}_0(0,W) - \Psi(Q_0)\right)^2\right] \\
&= uE_0\left[\frac{\tilde{Q}_0(1,W)\left(1 - u\tilde{Q}_0(1,W)\right)}{g_0(1|W)} + \frac{\tilde{Q}_0(0,W)\left(1 - u\tilde{Q}_0(0,W)\right)}{g_0(0|W)}\right] + u^2 E_0\left[\left(\tilde{Q}_0(1,W) - \tilde{Q}_0(0,W) - \Psi(\tilde{Q}_0)\right)^2\right]
\end{aligned}$$

where $\Psi(Q_0) = E_0[\bar{Q}_0(1,W) - \bar{Q}_0(0,W)]$ and $\Psi(\tilde{Q}_0) = E_0[\tilde{Q}_0(1,W) - \tilde{Q}_0(0,W)]$.

Now consider the outcome Y to be a proportion. Specifically, suppose the outcome Y is the average of k independent Bernoulli's with probability $\bar{Q}_0(A,W) \in [0,u]$. The variance of the efficient influence curve at the true but unknown distribution P_0 is then

$$\begin{aligned}
Var[D^*(Q_0, g_0)] =& \frac{u}{k} E_0\left[\frac{\tilde{Q}_0(1,W)\left(1 - u\tilde{Q}_0(1,W)\right)}{g_0(1|W)} + \frac{\tilde{Q}_0(0,W)\left(1 - u\tilde{Q}_0(0,W)\right)}{g_0(0|W)}\right] \\
&+ \left(\frac{u}{k}\right)^2 E_0\left[\left(\tilde{Q}_0(1,W) - \tilde{Q}_0(0,W) - \Psi(\tilde{Q}_0)\right)^2\right]
\end{aligned}$$

Appendix B: Asymptotic variance at a misspecified limit $\bar{Q}(A,W)$

Suppose we have n independent, identically distributed (i.i.d.) observations of $O = (W, A, Y)$ drawn from some P_0 in the semiparametric statistical model $\mathcal{M}$. Further suppose, for discussion, that the exposure mechanism $g_0(A|W)$ is known and satisfies the positivity assumption. Again, we consider Y to be binary for simplicity. Now consider two TMLEs $\Psi_{1,n}(P_n)$ and $\Psi_{2,n}(P_n)$, which are functions of the empirical distribution P_n. Suppose the first constrains the estimated risks $\bar{Q}_n(A,W)$ to be ≤ 1, while the second constrains its estimated risks $\bar{Q}_n(A,W)$ to be $\leq u$. In other words, $\Psi_{2,n}(P_n)$ ensures the estimated risks are within the model bounds. Since both TMLEs solve the efficient influence curve equation, they are double robust and will be consistent even if $\bar{Q}_n(A,W)$ converges to a misspecified limit. The second TMLE, however, will often be more efficient when $\bar{Q}_0(A,W)$ is inconsistently estimated.

Under regularity conditions, the asymptotic variance of the first estimator $\Psi_{1,n}(P_n)$ is given by

the variance of the efficient influence curve at the misspecified limit $\bar{Q}(A,W)$ divided by n:

$$\begin{aligned} nVar[\Psi_{1,n}(P_n)] = Var[D^*(P_0)(O)] &+ E_0\left[\frac{\big(1-g_0(1|W)\big)}{g_0(1|W)}\big(\bar{Q}_0(1,W)-\bar{Q}(1,W)\big)^2\right] \\ &+ E_0\left[\frac{\big(1-g_0(0|W)\big)}{g_0(0|W)}\big(\bar{Q}_0(0,W)-\bar{Q}(0,W)\big)^2\right] \\ &+ 2E_0\left[\big(\bar{Q}_0(1,W)-\bar{Q}(1,W)\big)\big(\bar{Q}_0(0,W)-\bar{Q}(0,W)\big)\right] \end{aligned}$$

where $Var[D^*(P_0)(O)]$ is the variance of the efficient influence curve at $P_0 \in \mathcal{M}$ as given above. The second and third terms, involving squared deviations between the true mean $\bar{Q}_0(A,W)$ and limit $\bar{Q}(A,W)$, are always positive. The last term, involving the product of these deviations, will be positive when both treatment-specific means are under-estimated or over-estimated. Thereby, when $\bar{Q}(A,W)$ approaches 1 for some treatment-covariate combinations, these terms will be positive and contribute substantially to the asymptotic variance of the first estimator.

In contrast, the asymptotic variance of the second estimator $\Psi_{2,n}(P_n)$ is

$$\begin{aligned} nVar[\Psi_{2,n}(P_n)] = Var[D^*(P_0)(O)] &+ u^2E_0\left[\frac{\big(1-g_0(1|W)\big)}{g_0(1|W)}\big(\tilde{Q}_0(1,W)-\tilde{Q}(1,W)\big)^2\right] \\ &+ u^2E_0\left[\frac{\big(1-g_0(0|W)\big)}{g_0(0|W)}\big(\tilde{Q}_0(0,W)-\tilde{Q}(0,W)\big)^2\right] \\ &+ 2uE_0\Big[\big(\tilde{Q}_0(1,W)-\tilde{Q}(1,W)\big)\big(\tilde{Q}_0(0,W)-\tilde{Q}(0,W)\big)\Big] \end{aligned}$$

where we have replaced the limit $\bar{Q}(A,W)$ with $u\tilde{Q}(A,W)$ for some $\tilde{Q}(A,W) \in [0,1]$. Since u is small, the contribution from misspecification of $\bar{Q}_0(A,W)$ is diminished in the second estimator, which enforces the constraints in the statistical model $\mathcal{M}$. Thereby, this estimator will be closer to achieving the efficiency bound even if $\bar{Q}_n(A,W)$ converges to a misspecified limit. This provides an asymptotic motivation for constructing a new TMLE, which guarantees the predicted probabilities are within model bounds and does not rely on them being nicely bounded by chance. Indeed, one could easily imagine a situation where an unconstrained estimator for $\bar{Q}_0(A,W)$ adds more covariates into the regression model as sample size increases and thereby converges to a limit outside the model bounds.

Finally, we note that if $\bar{Q}_n(A,W)$ converges to the true $\bar{Q}_0(A,W)$, then the two estimators $\Psi_{1,n}(P_n)$ and $\Psi_{2,n}(P_n)$ will be asymptotically equivalent. Both estimators will achieve the efficiency bound in that their asymptotic variance will be given by the variance of the efficient influence curve at P_0 divided by sample size. Their finite sample performance, however, is still likely to be different. The proof is available upon request.

Appendix C - Sample R Code

Below, we provide sample R code to minimize the log-likelihood loss for $\tilde{Y} = (Y-\ell)/(u-\ell)$. Full code for implementing the rare outcomes TMLE algorithm is available from the authors by request.

```
#############
# LogLikelihood() -
# function to calculate the loglikelihood loss for a bounded Y
#        with values possibly outside of [0,1]
# this uses on a parametric regression model for P(Y | A,W)
```

```
# input: beta (initial values for coefficients), outcome Y, design matrix X
# returns: negative log likelihood loss
#############
LogLikelihood<- function(beta, Y, X){
  pi<- plogis( X%*%beta )                 # P(Y|A,W)= expit(beta0 + beta1*X1+beta2*X2...)
  pi[pi==0] <- .Machine$double.neg.eps        # to prevent taking the log of 0
  pi[pi==1] <- 1-.Machine$double.neg.eps
  logLike<- sum( Y*log(pi)  + (1-Y)*log(1-pi)  )
  return(-logLike)
}
############
# grad- corresponding function to calculate the gradient
# other optimization routines (e.g. Nelder-Mead) do not use the gradient
######
grad<- function(beta, Y, X){
  pi<- plogis( X%*%beta )        # P(Y|A,W)= expit(beta0 + beta1*X1+beta2*X2...)
  pi[pi==0] <- .Machine$double.neg.eps        # for consistency with above
  pi[pi==1] <- 1-.Machine$double.neg.eps
  gr<- crossprod(X, Y-pi)        # gradient is -residual*covariates
  return(-gr)
}


# Example: Data generating experiment for Simulation 1
set.seed(123)
n=2500
W1<- rnorm(n, 0, .25)
W2<- runif(n, 0, 1)
W3<- rbinom(n, size=1, 0.5)
A<- rbinom(n, size=1, prob= plogis(-.5+ W1+W2+W3) )
pi<- plogis(-3+ 2*A + 1*W1+2*W2-4*W3 + .5*A*W1)/15
Y<- rbinom(n, size=1, prob= pi)
sum(Y)
# 29

# Qbounds (l,u)= (0,0.065)
l=0; u=0.065
#create the design matrix
X <- model.matrix(as.formula(Y~W1+W2+W3+A*W1))
# transform Y to Y.tilde in between (l,u)
Y.tilde<- (Y - l)/(u-l)
summary(Y.tilde)
# Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
# 0.0000  0.0000  0.0000  0.1785  0.0000 15.3800

# call to the optim function.
# par: initial parameter estimates; f:function to minimize; gr: gradient
# arguments to LogLikelihood() & grad() are Y and X
optim.out <- optim(par=rep(0, ncol(X)), fn=LogLikelihood, gr=grad,
        Y=Y.tilde, X=X, method="BFGS")
# see optim help files for more details and other optimization routines

# get parameter estimates
beta<- optim.out$par
# get predicted values and transform to proper scale
pred.prob.optim <- plogis(X%*%beta)*(u-l) + l
# compare with
pred.prob.glm<- predict(glm(Y~W1+W2+W3+A*W1, family="binomial"), type="response")
predictions<- data.frame(optim=pred.prob.optim,  glm=pred.prob.glm)
```

```
summary(predictions*100)
#            optim               glm
# Min.   :0.00001   Min.   : 0.000364
# 1st Qu.:0.02923   1st Qu.: 0.070958
# Median :0.14395   Median : 0.159921
# Mean   :1.16000   Mean   : 1.160000
# 3rd Qu.:1.62741   3rd Qu.: 1.678541
# Max.   :6.30201   Max.   :10.321924
```

6 Acknowledgements

The authors would like to thank Drs. Jennifer Ahern, Susan Gruber, Sam Lendle and Jeremy Taylor for their helpful comments and suggestions. This work was supported, in part, by the National Institutes of Health under award number R01 AI074345. The content is solely the responsibility of the authors and does not necessarily represent the officials views of the National Institutes of Health.

References

A. Abadie and G.W. Imbens. Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74:235–267, 2006.

J.A. Anderson. Separate sample logistic discrimination. *Biometrika*, 59:19–35, 1972.

N. Beck, G. King, and L. Zeng. Improving quantitative studies of international conflict: A conjecture. *American Political Science Review*, 94(1):21–25, 2000.

P.J. Bickel, C.A.J. Klaassen, Y. Ritov, and J.A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press, Baltimore, 1993.

Birthplace in England Collaborative Group. Perinatal and maternal outcomes by planned place of birth for healthy women with low risk pregnancies: the Birthplace in England national prospective cohort study. *British Medical Journal*, 343(7840):d7400, 2011.

L.E. Braitman and P.R. Rosenbaum. Rare outcomes, common treatments: Analytic strategies using propensity scores. *Annals of Internal Medicine*, 137(8):693–696, 2002.

M.S. Cepeda, R. Boston, J.T. Farrar, and B.L. Strom. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *American Journal of Epidemiology*, 158(3):280–287, 2003.

J. Concato, A.R. Feinstein, and T.R. Holford. The risk of determining risk with multivariable models. *Annals of Internal Medicine*, 118:201–210, 1993.

S. Gruber and M.J. van der Laan. A targeted maximum likelihood estimator of a causal effect on a bounded continuous outcome. *The International Journal of Biostatistics*, 6(1):Article 26, 2010. doi: 10.2202/1557-4679.1260.

S. Gruber and M.J. van der Laan. An application of targeted maximum likelihood estimation to the meta-analysis of safety data. *Biometrics*, 69:254–262, 2013. doi: 10.1111/j.1541-0420.2012.01829.x.

F.E. Harrell, Jr., K.L. Lee, and D.B. Mark. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15:361–387, 1996.

M.A. Hernán, B. Brumback, and J.M. Robins. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology*, 11(5):561–570, 2000. PMID: 10955409.

M.M. Joffe and P.R. Rosenbaum. Invited commentary: propensity scores. *American Journal of Epidemiology*, 150(4):327–333, 1999.

G. King and L. Zeng. Logistic regression in rare events data. *Political Analysis*, 9(2):137–163, 2001.

S.D. Lendle, B. Fireman, and M.J. van der Laan. Targeted maximum likelihood estimation in safety analysis. *J Clin Epidemiol*, 66:S91–S98, 2013.

R. Mansson, M.M. Joffe, W. Sun, and S. Hennessy. On the estimation and use of propensity scores in case-control and case-cohort studies. *American Journal of Epidemiology*, 166(3):332–339, 2007.

J. Neyman. Sur les applications de la theorie des probabilites aux experiences agricoles: Essai des principes (In Polish). English translation by D.M. Dabrowska and T.P. Speed (1990). *Statistical Science*, 5:465–480, 1923.

E. Patorno, R.J. Glynn, S. Hernández-Díaz, J. Liu, and S. Schneeweiss. Studies with many covariates and few outcomes: selecting covariates and implementing propensity-score-based confounding adjustments. *Epidemiology*, 26(2):268–278, 2014. doi: 10.1097/EDE.0000000000000069.

J. Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, New York, 2000. second ed., 2009.

P. Peduzzi, J. Concato, E. Kemper, T.R. Holford, and A.R. Feinstein. A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*, 49(12):1373–1379, 1996.

M.L. Petersen and M.J. van der Laan. Causal models and learning from data: Integrating causal modeling and statistical estimation. *Epidemiology*, 25(3):418–426, 2014.

M.L. Petersen, K.E. Porter, S. Gruber, Y. Wang, and M.J. van der Laan. Diagnosing and responding to violations in the positivity assumption. *Statistical Methods in Medical Research*, 21(1):31–54, 2012. doi: 10.1177/0962280210386207.

R.L. Prentice and N.E. Breslow. Retrospective studies and failure time models. *Biometrika*, 65(1): 153–158, 1978.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. URL http://www.R-project.org.

RARECARENet. Information network on rare cancers, 2014. URL http://www.rarecarenet.eu/rarecarenet/.

J.M. Robins. A new approach to causal inference in mortality studies with sustained exposure periods–application to control of the healthy worker survivor effect. *Math Mod*, 7:1393–1512, 1986.

J.M. Robins. [Choice as an alternative to control in observational studies]: Comment. *Statistical Science*, 14(3):281–293, 1999.

J.M. Robins. Robust estimation in sequentially ignorable missing data and causal inference models. In *1999 Proceedings of the American Statistical Association*, pages 6–10, Alexandria, VA, 2000. American Statistical Association.

P.R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55, 1983.

D.B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974. doi: 10.1037/h0037350.

J.S. Sekhon. Multivariate and propensity score matching software with automated balance optimization: the matching package for R. *Journal of Statistical Software*, 42(7):1–52, 2011.

M. van der Laan and S. Rose. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer, Berlin Heidelberg New York, 2011.

M.J. van der Laan. Estimation based on case-control designs with known prevalence probability. *The International Journal of Biostatistics*, 4(1):Article 17, 2008.

M.J. van der Laan and J.M. Robins. *Unified Methods for Censored Longitudinal Data and Causality*. Springer-Verlag, New York Berlin Heidelberg, 2003.

M.J. van der Laan and D.B. Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1):Article 11, 2006. doi: 10.2202/1557-4679.1043.

M.J. van der Laan, E.C. Polley, and A.E. Hubbard. Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1):Article 25, 2007.

World Health Organization. Global tuberculosis report 2013. Geneva, 2013.